SDC 19

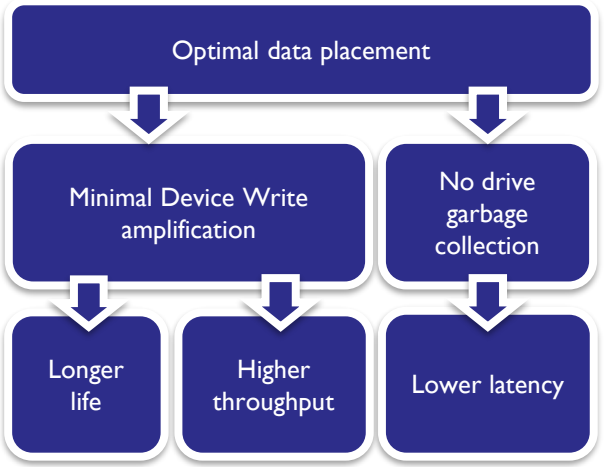
September 23-26, 2019
Santa Clara, CA

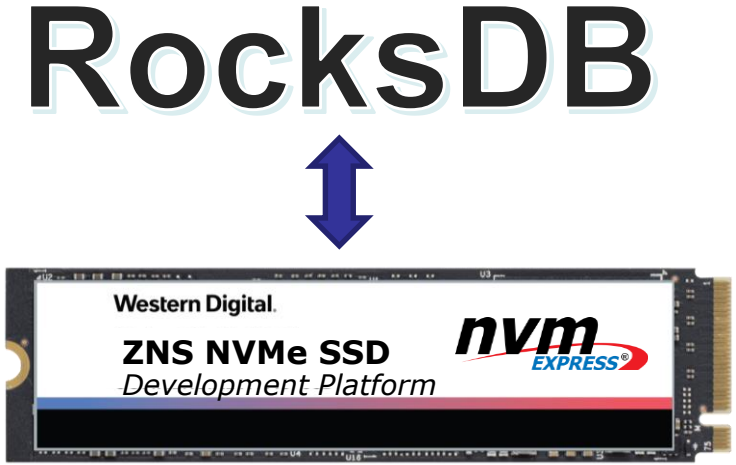
Accelerating RocksDB with NVMe™ Zoned SSDs

Hans Holmberg, R&D Technologist
Emerging System Architectures Group

Western Digital

RocksDB on Zoned NVMe™ SSDs





2019 Storage Developer Conference. © Western Digital Corporation or its affiliates. All Rights Reserved.

Agenda

- Zoned Namespaces 101
- Adapting RocksDB for Zoned SSDs
- Demo
- Results
- What's next?

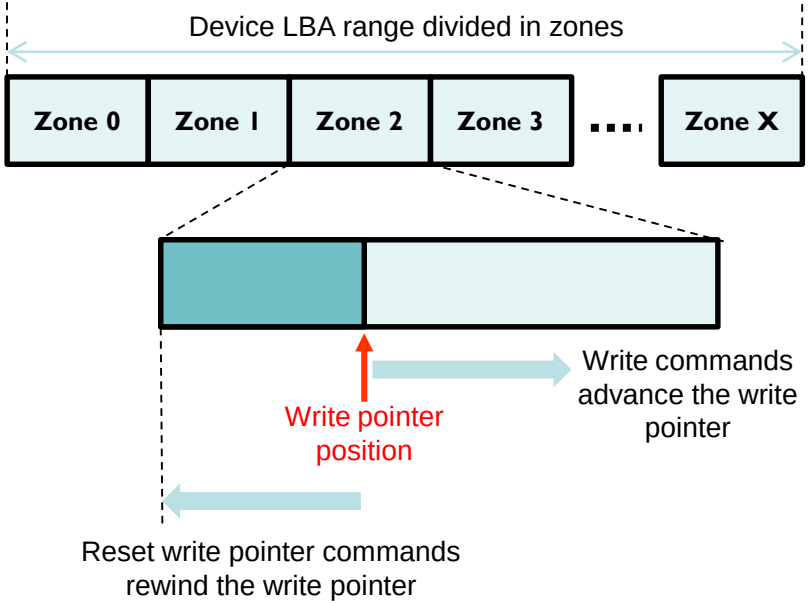
2019 Storage Developer Conference. © Western Digital Corporation or its affiliates. All Rights Reserved.

Zoned Namespaces 101

What are Zoned Block Devices?

The new paradigm in storage

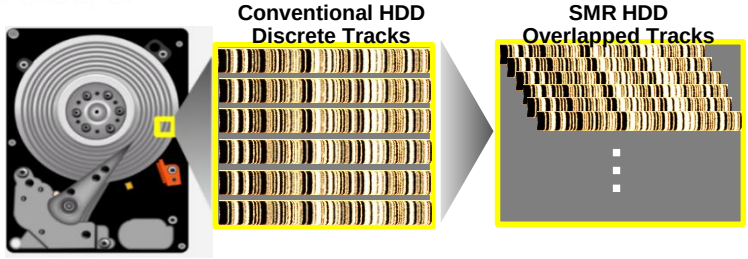
- The storage device logical block addresses are divided into ranges of zones.
- Writes within a zone must be sequential.
- The zone must be erased before it can be rewritten.



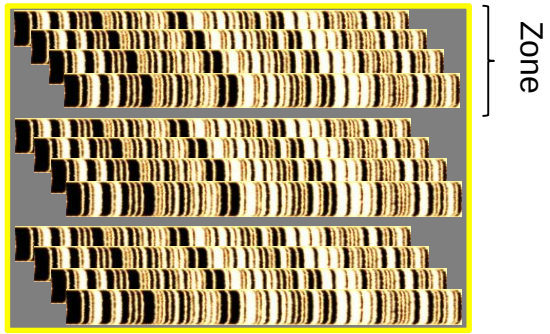
2019 Storage Developer Conference. © Western Digital Corporation or its affiliates. All Rights Reserved.

Zoned Storage on SMR

- SMR (Shingled Magnetic Recording)
  - Enables areal density growth
  - Shares flash access model
    - Erase before re-write



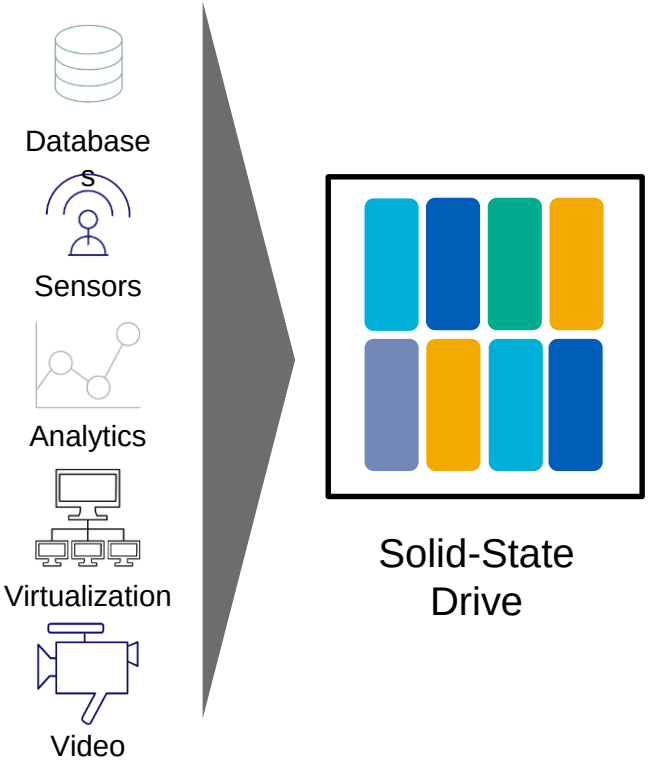
- Zoned Access
  - Zoned Block I/F standardized in INCITS
    - Zoned Block Commands (ZBC): SAS
    - Zoned ATA Commands (ZAC): SATA
  - Host/Device cooperate to optimize RMW aspect of SMR by enforcing sequential writes and enabling host FTL model



2019 Storage Developer Conference. © Western Digital Corporation or its affiliates. All Rights Reserved.

Ubiquitous Workloads

The cloud applies multiple workloads to a single SSD



SSDs write log-structured to the media that requires garbage collection

Multiplex data streams onto the same garbage collection units

Increases

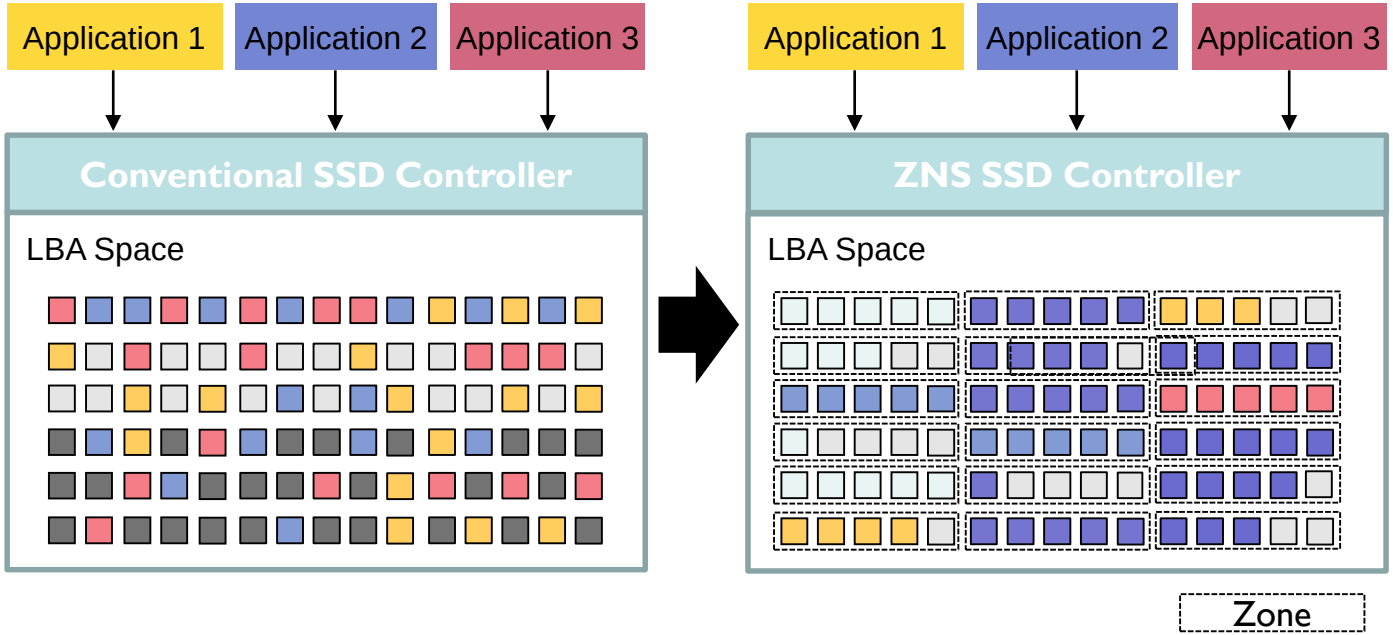
Write Amplification, Over-Provisioning and thereby Cost

Decreases

throughput and latency predictability

2019 Storage Developer Conference. © Western Digital Corporation or its affiliates. All Rights Reserved.

Zones for Solid State Drives



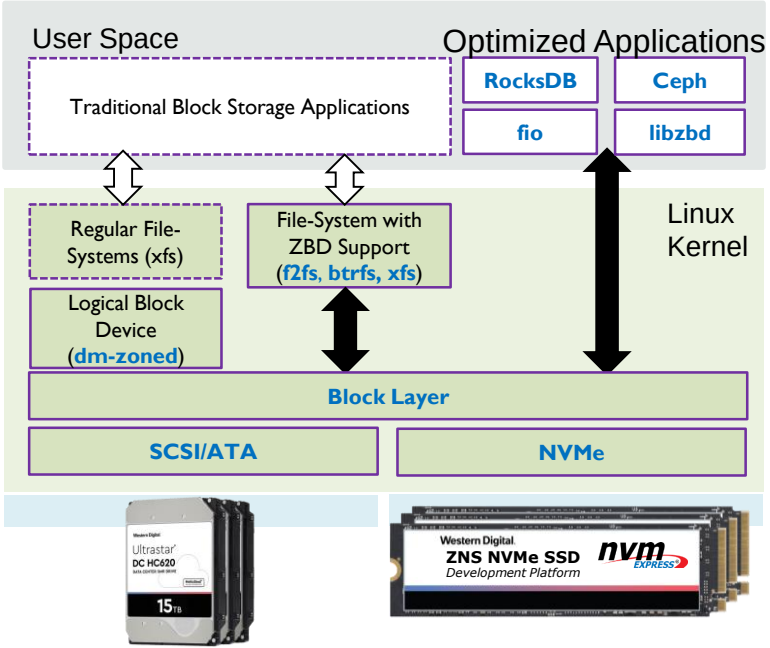
Eliminate data streams multiplexing:

- **Significantly decreases** write amplification, over-provisioning and **thereby reduces cost**
- **Increases throughput and latency predictability**

2019 Storage Developer Conference. © Western Digital Corporation or its affiliates. All Rights Reserved.

ZNS: Synergies w/ ZAC/ZBC software ecosystem

- Device exposed as a Zoned Block Device (ZBD)
- Reuse existing work already done for ZAC/ZBC devices
- Existing ZBD-aware file systems & device mappers “just work”
  - Few additions to support to ZNS
- Integrates with file-systems and applications
  - RocksDB, Ceph, fio, libzbd, …
- ZAC/ZBC devices are already in production at technology adopters and a mature storage stack is available through the Linux® eco-system



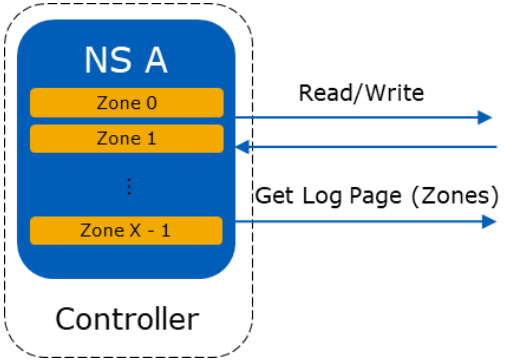
***= Enhanced data paths for SMR/ZNS drives**

2019 Storage Developer Conference. © Western Digital Corporation or its affiliates. All Rights Reserved.

Zoned Namespaces

- **Ongoing Technical Proposal in the NVMe working group**
- New Zoned Command Set – Inherits the NVM Command Set and adds zone support.
- Aligns to the existing host-managed models defined in the ZAC/ZBC specifications.
  - Note that it does not map 1:1. Beware of the details.
- Optimized for Solid State Drives
  - Zone Capacity
  - Zone Append
  - Zone Descriptors

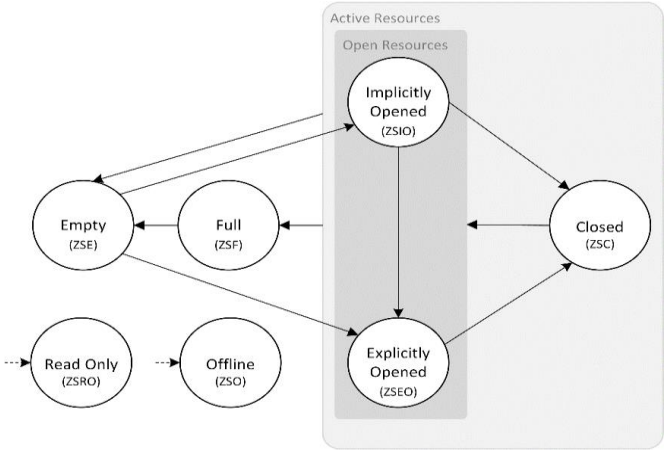
Under review

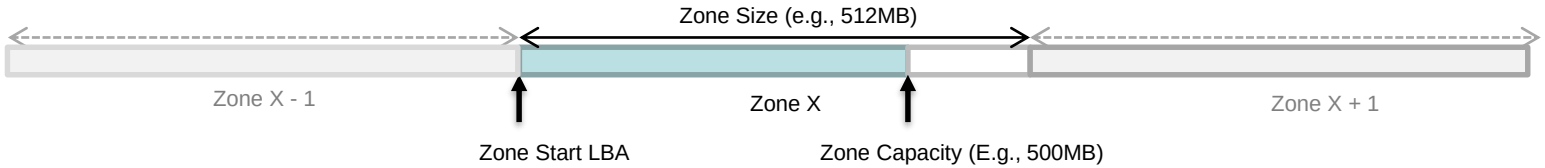


2019 Storage Developer Conference. © Western Digital Corporation or its affiliates. All Rights Reserved.

Host-Managed Zoned Block Devices

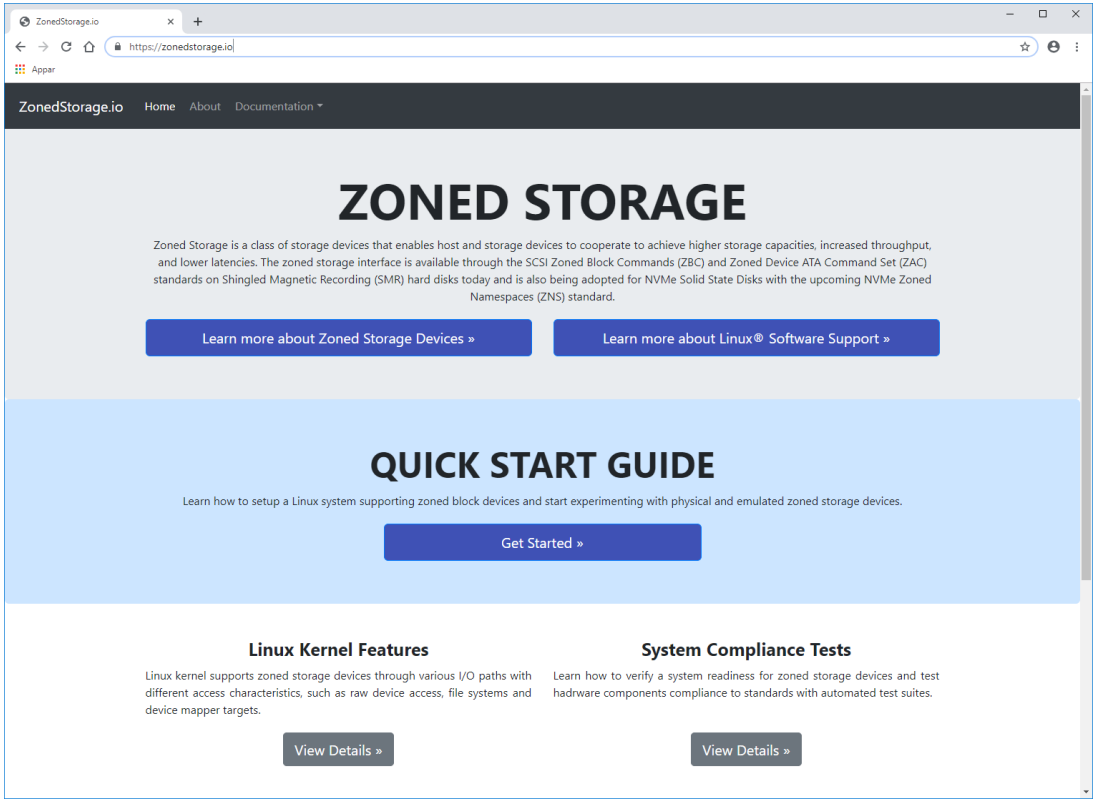
- Zone States
  - Empty, Implicitly Opened, Explicitly Opened, Closed, Full, Read Only, and Offline.
  - Changes state upon writes, zone management commands, and device resets.
- Zone Management
  - Open Zone, Close Zone, Finish Zone, and Reset Zone
- Zone Size & Zone Capacity(NEW)
  - Zone Size is fixed
  - Zone Capacity is the writeable area within a zone





2019 Storage Developer Conference. © Western Digital Corporation or its affiliates. All Rights Reserved.

ZonedStorage.IO



2019 Storage Developer Conference. © Western Digital Corporation or its affiliates. All Rights Reserved.

RocksDB on ZNS

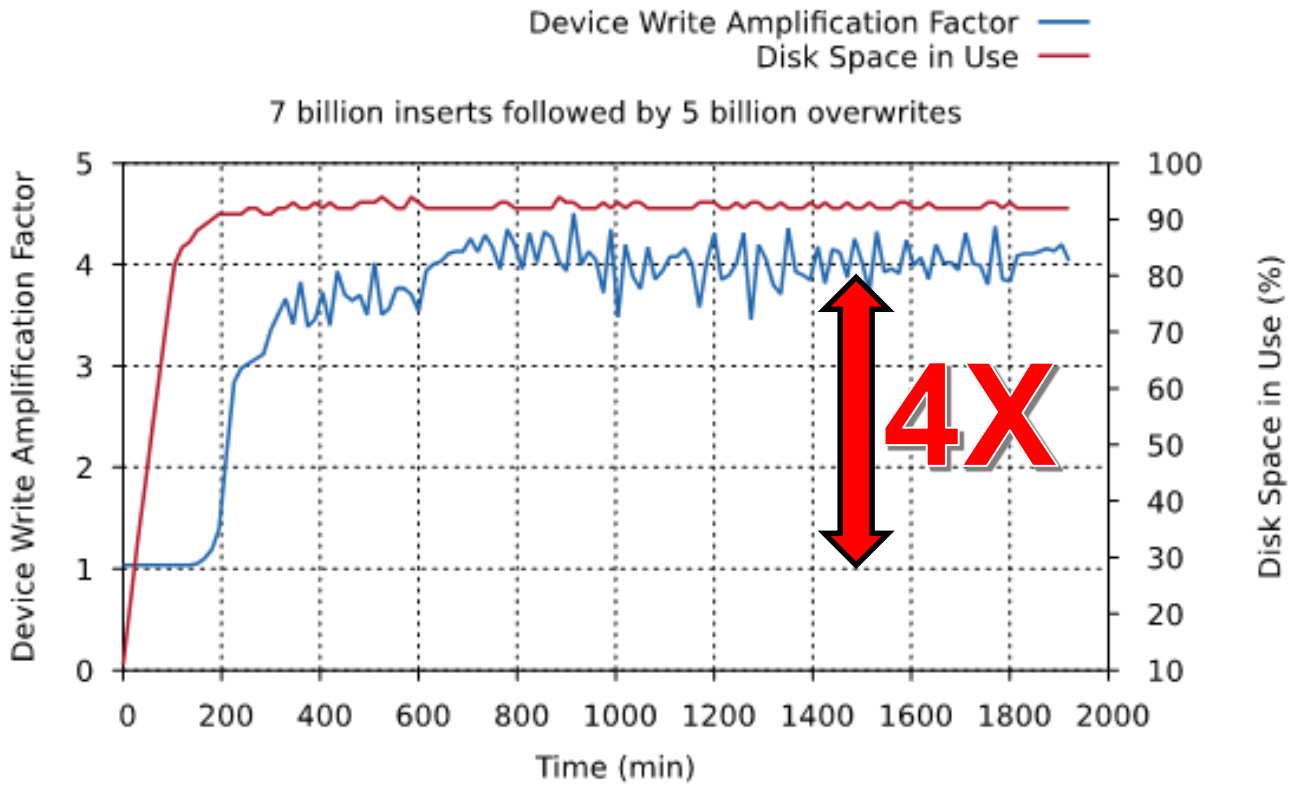
RocksDB, a good fit for ZNS

- Persistent key-value store for fast storage environments
- Log-structured, flash friendly
- Customizable storage back ends

RocksDB

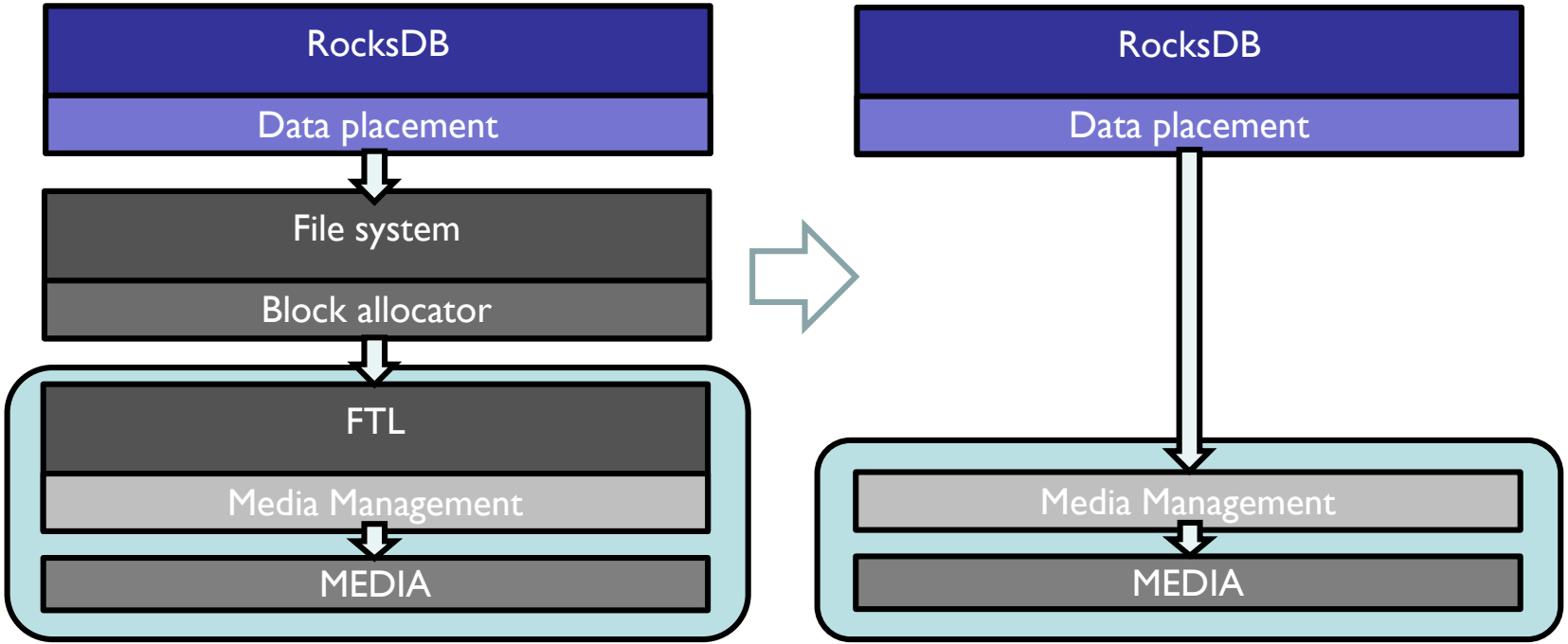
2019 Storage Developer Conference. © Western Digital Corporation or its affiliates. All Rights Reserved.

WA on a conventional SSD



2019 Storage Developer Conference. © Western Digital Corporation or its affiliates. All Rights Reserved.

Target: End-to-end-integration



2019 Storage Developer Conference. © Western Digital Corporation or its affiliates. All Rights Reserved.

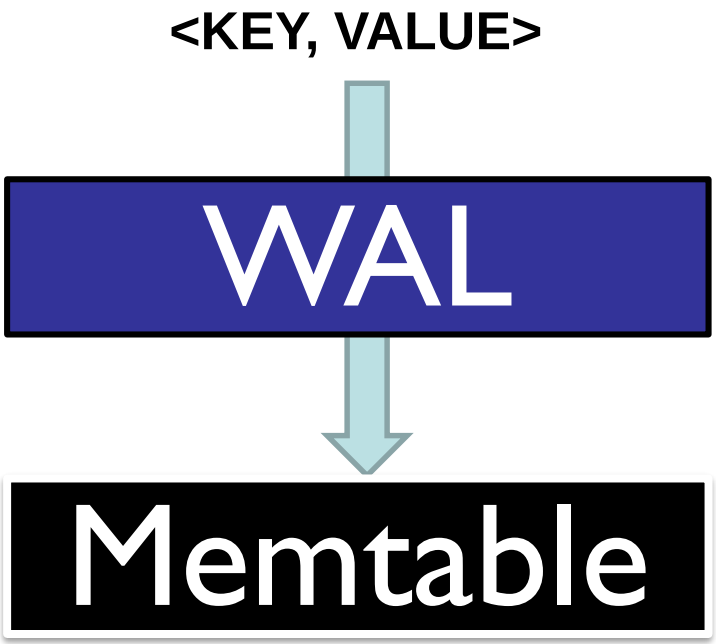
Challenges

- Multiple, parallel files being written to
  - Map each file to a set of zones
- All writes must be sequential and ordered
  - Use direct I/O and the deadline scheduler
- Limits on number of open zones
  - Finish zones when done with writes

2019 Storage Developer Conference. © Western Digital Corporation or its affiliates. All Rights Reserved.

RocksDB on-disk data structures

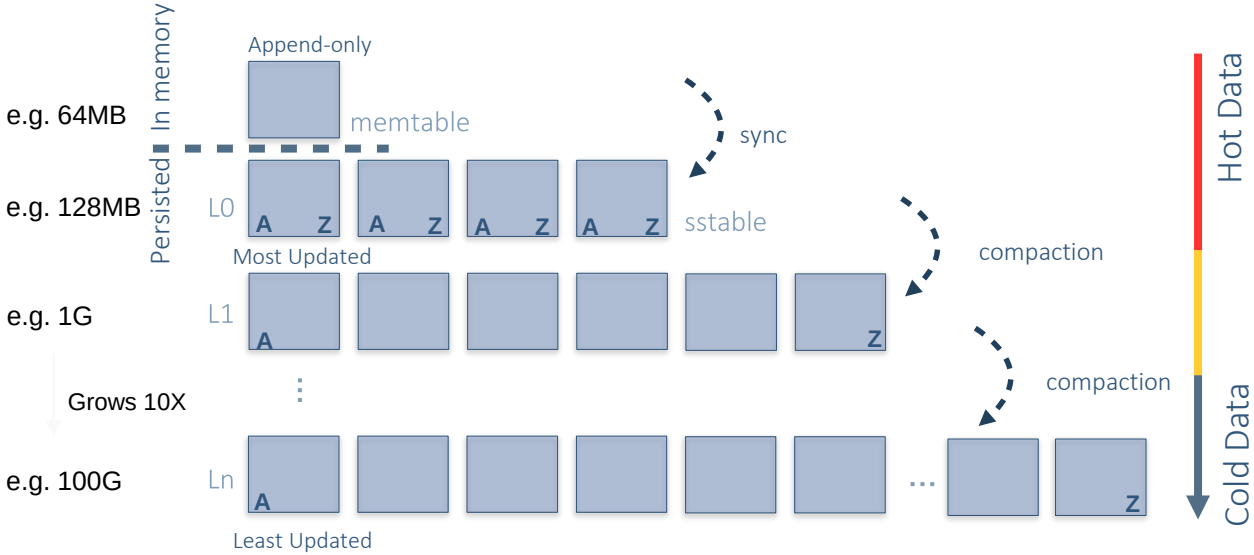
Writeahed-log (WAL)



2019 Storage Developer Conference. © Western Digital Corporation or its affiliates. All Rights Reserved.

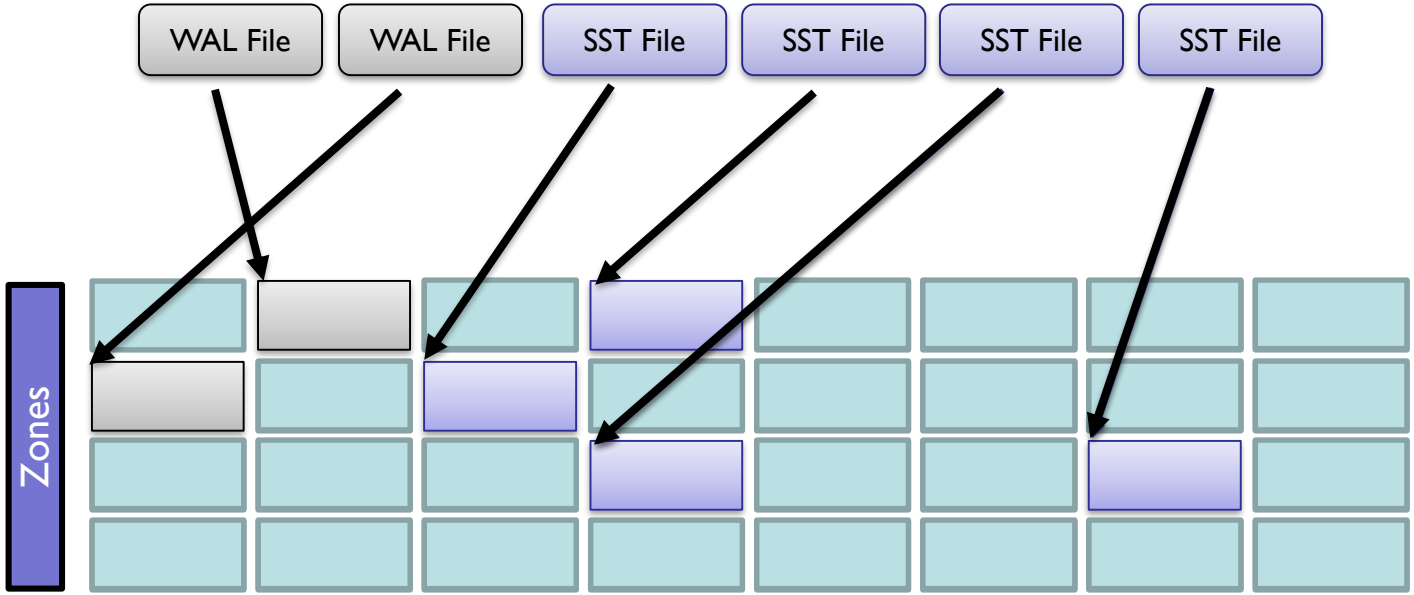
RocksDB on-disk data structures

Sorted string tables



2019 Storage Developer Conference. © Western Digital Corporation or its affiliates. All Rights Reserved.

Mapping files to zones



2019 Storage Developer Conference. © Western Digital Corporation or its affiliates. All Rights Reserved.

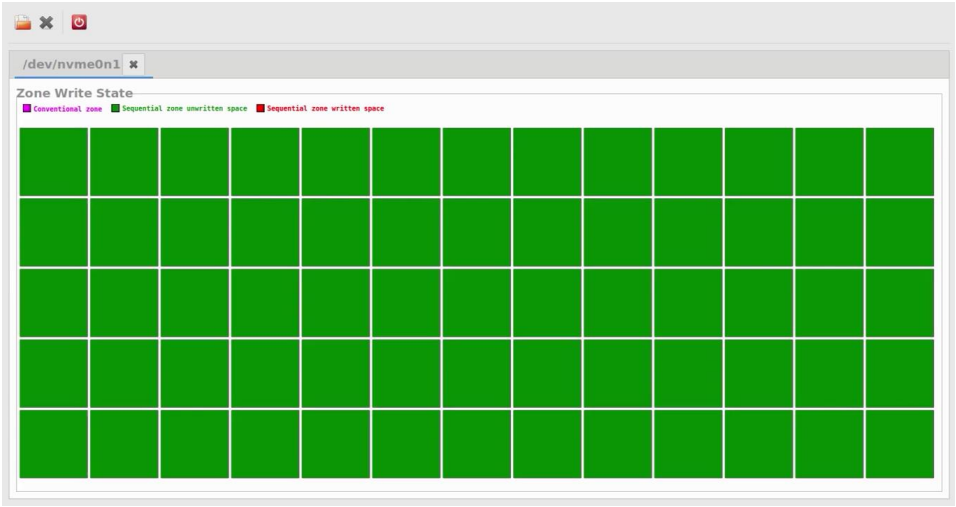
Approach

- Files are mapped to zones
- Zone management through file management
  - Zones are allocated when creating a new file
  - Zones are released after file deletion
  - Zones can be rewritten after being reset

No device-side garbage collection

2019 Storage Developer Conference. © Western Digital Corporation or its affiliates. All Rights Reserved.

Demo (Random Insert Workload)



2019 Storage Developer Conference. © Western Digital Corporation or its affiliates. All Rights Reserved.

Results

Smart data placement	**~1X device write amplification** 3-6X WA's measured on conventional drives (28% OP)
No on-drive garbage collection	**20% increase in capacity** Compared to a conventional 28% OP SSD
	20% TCO reduction Increases lifetime/writes significantly

2019 Storage Developer Conference. © Western Digital Corporation or its affiliates. All Rights Reserved.

Conclusions

- Easy to leverage flash-friendly data placement
  - ZNS enables applications to become flash-optimal
- Zoned Block Device Software Eco-system already available
  - Libraries, tools, emulation
- Easy to integrate with existing storage stack

2019 Storage Developer Conference. © Western Digital Corporation or its affiliates. All Rights Reserved.

What's next?

- Upstream support to RocksDB
- More ZNS end-to-end-integration:
  - Databases (LSM-based, logs, ...)
  - Filesystems (btrfs, ceph, ...)
  - Cloud infrastructure

2019 Storage Developer Conference. © Western Digital Corporation or its affiliates. All Rights Reserved.

SDC 19

September 23-26, 2019
Santa Clara, CA

Thanks!

Western Digital and the Western Digital logo are registered trademarks or trademarks of Western Digital Corporation or its affiliates in the US and/or other countries. The NVMe word mark and NVM Express design mark are trademarks of NVM Express, Inc. All other marks are the property of their respective owners.